DOCUMENT RESUME

ED 277 734 TM 870 010

AUTHOR Lenel, Julia C.; Gilmer, Jerry S.
TITLE The Effect of Keying All Options Correct on Equating Functions and Scores.
PUB DATE Apr 86
NOTE 16p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Equated Scores; *Item Analysis; Latent Trait Theory; Licensing Examinations (Professions); Multiple Choice Tests; Postsecondary Education; Scores; *Scoring; *Scoring Formulas; Testing Problems; Test Items; Test Length; Test Reliability
IDENTIFIERS *Linear Equating Method

ABSTRACT

In some testing programs an early item analysis is performed before final scoring in order to validate the intended keys. As a result, some items which are flawed and do not discriminate well may be keyed so as to give credit to examinees no matter which answer was chosen. This is referred to as allkeying. This research examined how varying the numbers of allkeyed items affects the equating function and resulting equated scores. The experimental conditions consisted of allkeying 0, 4, 10, and 25 items. The examination was a 200-item multiple choice licensing examination. Over 3,500 examinee records were studied. The results showed virtually no differences in scaled score means across the experimental conditions. Although the equating procedures compensated for the changes that occurred as more items were allkeyed, the effect of allkeying on an individual's scaled score will depend on the individual's performance on the allkeyed items. The results suggest that an item should not be allkeyed unless it is clear that there is no defensible answer among the options. (Author/GDC)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED277734

# The Effect of Keying All Options Correct on Equating Functions and Scores

Julia C. Lenel
Jerry S. Gilmer

The American College Testing Program

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. Lenel

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the Annual Meeting

of the American Educational Research Association

San Francisco, April 1986

TM 870 010

The Effect of Keying All Options Correct

on Equating Functions and Scores

ABSTRACT

In some testing programs an "early item analysis" is performed before final scoring in order to validate the intended keys. As a result, some items may be keyed so as to give credit to examinees no matter which answer was chosen. (This is referred to as allkeying in this paper.) The purpose of this research is to examine how varying the numbers of allkeyed items affects the equating function and resulting equated scores. The experimental conditions consisted of allkeying zero, four, ten, and twenty-five items. The results showed virtually no differences in scaled score means across the experimental conditions. Although the equating procedures compensated for the changes that occurred as more items were allkeyed, the effect of allkeying on an individual's scaled score will depend on the individual's performance on the allkeyed items. The results suggest that an item should not be allkeyed unless it is clear that there is no defensible answer among the options.

## INTRODUCTION

In many standardized testing programs, an "early item analysis" is performed before final scoring in order to ensure the quality and fairness of the items in the examination. The purpose of this item analysis is to identify items that performed poorly, as indicated by indices of difficulty and discrimination. If a review of the item content reveals that an item is flawed, the item may be scored all options correct (hereafter referred to as allkeying). The number of allkeyed items varies from one test form to another.

Although the practice of allkeying items prior to equating and final scoring is not uncommon in standardized testing, the effects of allkeying on equating functions and equated scores have received little attention in the literature. Dorans (1983) examined the effect of deleting an item (i.e., scoring an item either all options correct or no options correct) on equating/scaling functions when IRT equating procedures were used. He found that the effect of deleting an item was dependent on the characteristics of the deleted item (i.e., difficulty, discrimination, and lower assymptote of the item characteristic curve) and the scoring method used (i.e., no options vs. all options correct). Dorans also found that when a flawed item was discovered after the equating process was completed, the change in scaled scores was much smaller when a new equating function was determined than when the item was simply rescored either no options correct or all options correct.

Dorans was concerned with the effect on IRT true score equating of allkeying (deleting) a single item that was identified as flawed only after equating and final scoring had occurred. The present study investigates the

effect on linear equating of allkeying several items that have been identified as flawed, based on statistical indices, before equating and final scoring. The purpose is to examine how varying the number of allkeyed items affects the equating function and individual scaled scores.

## METHOD

The data in this study are from a nationally administered licensure examination administered to more than 18,000 candidates. A spaced sample of 3,588 examinee records was selected for this study. The examination is composed of 200 multiple-choice items. Each item is classified into one of six content areas. Two of the content areas each contain 20 percent of the total items while the other four content areas each contain 15 percent of the total items. Forty of the items were chosen from a previous test form and constitute an internal anchor that is used to equate scores on the current form to a standard score scale. The equators were chosen to be both statistically and content-representative of the complete form from which they were chosen.

In this study, the experimental conditions consisted of allkeying zero, four, ten, or twenty-five items. Although scoring twenty-five items all options correct rarely occurs in practice, this condition was included for theoretical interest. The specific items chosen for the four allkeying conditions were among the items flagged during the early item analysis as statistically questionable. None of the allkeyed items was an equator. In this study, flawed items were scored all options correct because that is standard practice on this licensure exam. Another option would be to score no options correct. If raw scores are equated, the choice of scoring method is arbitrary. Items scored all options or no options correct have no

differential psychometric impact on examinee scores. Either scoring method effectively deletes the items from the test, and the results, after equating, are identical.

A second consideration in choosing groups of items to be allkeyed was the distribution of the items across content areas. When the equators were selected for this exam, they were chosen to reflect the percentages of items in each content area. However, the allkeying of items upsets the balance between the equators and the full test. Klein and Jarjoura (1985) found that anchors that were not representative of the test as a whole resulted in inaccurate equating. Representativeness was defined in terms of the distribution of items across content areas. In a representative anchor the percentage of equating items in each content area reflects the percentage of items in each content area for the full exam.

In view of this finding, an attempt was made to balance the allkeyed items across the six content areas. However, this was not entirely possible because some content areas had very few flagged items. The number of allkeyed items in each content area for the four conditions is listed in Table 1. Because complete balancing was not possible, a small degree of nonrepresentativeness was introduced. In general, the items that are allkeyed as a result of an early item analysis are distributed across content areas and thus introduce only slight nonrepresentativeness. However, there are other rekeying situations that could have a more serious effect on the representativeness of equators. For example, if a group of items from a single content area (e.g. a multi-item set) was allkeyed, the balance between the percentage of equators and total items in that content area would be upset. The findings of Klein and Jarjoura suggest that such an occurrence could affect the accuracy of equating. In order to test this hypothesis, a

fifth condition was added in which all the allkeyed items were chosen from a single content area. Two multi-item sets, one containing four items and the other five items, were allkeyed. The nine allkeyed items represented 23 percent of the total number of items in that content area. Again, none of the nine items was an equator. The statistical characteristics of the allkeyed items in the single content area condition (SC) differed somewhat from the other allkeyed items. These items were not originally flagged as statistically questionable and therefore tended to have higher difficulty values and higher indices of discrimination. Table 1 also presents the average item difficulty and discrimination of the allkeyed items for each of the five conditions.

All 3,588 records were scored under each of the five allkeying conditions. Following rescoring, equating functions were derived for each condition using two linear equating procedures: the Tucker method and the Levine equally reliable method. These methods were chosen because they are the methods generally employed in equating the examination used in this study. After the equating functions had been derived, basic summary statistics were obtained as well as raw and equated scores at specific points on the score scale.

## RESULTS

The raw score means and standard deviations, equated score means and standard deviations, and slopes and intercepts for the two linear equating procedures and five allkeying conditions are shown in Table 2. As expected, raw score means increased as greater numbers of items were allkeyed. For the two extreme conditions of zero and 25 allkeyed items, the raw score means were 123.537 and 140.108, respectively. However, for both methods of equating, equated score means remained virtually unchanged across the five conditions.

The largest difference in scaled score means was less than .03. It is clear that, with regard to mean scores, the equating procedures were successful in compensating for the changes in difficulty that occurred as more items were allkeyed.

The compensatory effect is more obvious if a comparison is made of the scaled scores that would be obtained for the same raw score in two different allkeying conditions. A raw score of 130 converts to a scaled score of 144 (using the Tucker method) if no items are allkeyed. When ten items are allkeyed, a raw score of 130 converts to a scaled score of 138. This difference occurs because the latter test is easier. To receive a scaled score of 144 on the test in which ten items were allkeyed, an examinee would need to obtain a raw score of 137.

Although the equating procedures do compensate, on the average, for changes introduced by allkeyed items, it is informative to look at the effect of allkeying on individual examinees. Table 3 shows the effect of allkeying items on the equated scores of two hypothetical examinees. It was assumed that both examinees obtained raw scores of 120 when no items were allkeyed. It was further assumed that examinee A chose the original key on every item that was later allkeyed, while examinee B chose a response other than the original key on the allkeyed items. The table shows that both examinees would obtain an equated score of 135 if no items were allkeyed. At the extreme of twenty-five allkeyed items, examinee A would receive an equated score of 118 rather than the original 135, while examinee B would receive an equated score of 143. This outcome is appropriate if the items were allkeyed because there was no correct response. In that case, an examinee who chose the original "correct" key would deserve no more credit than an examinee who chose an "incorrect" response. However, if an item were allkeyed because it did not

work statistically, i.e., had very low indices of difficulty and discrimination, but still had a justifiably correct answer, the result would be to penalize those examinees who knew that answer and reward those who did not.

A comparison of the zero allkeying condition and the single content area condition in Table 2 shows very little difference in scaled score means and standard deviations. The scaled score means for these two conditions differed by less than .01.

## DISCUSSION

The purpose of this study was to assess the effect of allkeying items on the linear equating function and individual scaled scores. The results showed that even in the extreme condition in which 25 items were allkeyed, the linear equating procedures are sufficiently robust to compensate for the changes introduced by allkeying. Although from a practical point of view it is encouraging to find that equating "works", it is somewhat surprising to find no effect even for the most extreme condition. One reason that allkeying these items may have had little effect on equating is that these items were contributing little to the test in the first place. Some support for this hypothesis can be found by examining the summary statistics in Table 1 and the KR-20 reliability coefficients for the examination under the four original allkeying conditions. The summary statistics show that the mean point biserial for the allkeyed items in all four conditions is less than .10. These items clearly do not discriminate between the good and poor examinees. The KR-20 values for the 0, 4, 10 and 25 item allkeying conditions were .865, .867, .869, and .870, respectively. Although the test effectively gets shorter as more items are scored all options correct, the reliability coefficients increase. This increase may indicate that the allkeyed items

were measuring something different than the other items and thus introduced noise into the measurement process. In any case, the change in reliability from 0 allkeyed items to 25 allkeyed items is quite small, only .005.

Although the allkeying of items had little effect on the mean scaled scores, it would be a mistake to conclude that allkeying had no effect. As the results in Table 3 show, the effect of allkeying items on an individual's scaled score depends on the individual's original response choice. Allkeying results in a decrease in scaled scores (relative to no allkeys) for individuals who chose the original "correct" response and in an increase in scaled scores for individuals who chose a response other than the key. The decrease in scaled scores for individuals choosing the "correct" response outcome can only be justified if the allkeyed items truly have no correct answer and the original key is no more correct than any of the distractors. A decision to allkey should be based on a consideration of the item content, not solely on the item statistics.

The failure to find an effect of allkeying on scaled scored means in the single content area allkeying condition is inconsistent with the Klein and Jarjoura study. Klein and Jarjoura found that non-representative anchors resulted in inaccurate equating. A partial explanation for this contradiction may be found in the degree of non-representativeness of the anchor forms. The percentages of equators for the SC condition in this study and for Klein and Jarjoura's nonrepresentative anchors are shown in Table 4 along with the percentage of items in the total exams used in each study. An examination of the table shows that there is a poorer match between percentages of equators and total items in both of the anchors used in the Klein and Jarjoura study than in the current study. It seems likely that the degree of non-representativeness of the anchor in the current study was not great enough to

affect the equating process. It should also be noted that Klein and Jarjoura manipulated the match by varying the number of equators chosen from each content area, while in the current study the lack of match was due to the allkeying of non-equating items. Although both manipulations result in a lack of match between percentages of equators and total test items, the effect on the equating process may not be the same.

In summary, this research examined the effect on linear equating of allkeying test items in a national standardized licensure testing program. The results showed that mean scaled scores remained virtually unchanged over all allkeying conditions examined. The linear equating procedures were sufficiently robust to withstand the violations of equating assumptions introduced by the manipulations in this study. However, the allkeying of flawed items can affect individual scaled scores and should be considered only after an analysis of the item content has revealed that no justifiably correct response appears among the options.

Table 1

Number of Allkeyed Items Across
Content Areas and Item Summary Statistics

| Allkeying Condition | Content Area 1 | 2 | 3 | 4 | 5 | 6 | Mean Diff. | Mean Pbis | KR20 |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | .00 | .00 | .865 |
| 4 | 1 | 0 | 0 | 0 | 1 | 2 | .26 | -.01 | .867 |
| 10 | 3 | 0 | 0 | 1 | 2 | 3 | .29 | .03 | .869 |
| 25 | 6 | 6 | 0 | 4 | 5 | 4 | .32 | .08 | .870 |
| SC | 9 | 0 | 0 | 0 | 0 | 0 | .66 | .18 | .860 |

Table 2

Raw and Scaled Score Means and Standard Deviations and Conversion Parameters for Five Allkeying Conditions and Two Methods of Linear Equating

| | | Number of Allkeyed Items | | | | |
|---|---|---|---|---|---|---|
| | | 0 | 4 | 10 | 25 | SC(9) |
| | Raw Score Mean | 123.537 | 126.079 | 130.069 | 140.108 | 126.642 |
| | Raw Score S.D. | 16.897 | 16.918 | 16.804 | 16.172 | 16.236 |
| | Scaled Score Mean | 137.878 | 137.876 | 137.871 | 137.860 | 137.873 |
| Tucker | Scaled Scores S.D. | 15.895 | 15.895 | 15.894 | 15.893 | 15.895 |
| Method | Slope | .941 | .940 | .946 | .983 | .979 |
| | Intercept | 21.664 | 19.422 | 14.844 | .166 | 13.893 |
| | Scaled Score Mean | 136.687 | 136.690 | 136.696 | 136.709 | 136.694 |
| Levine | Scaled Score S.D. | 15.847 | 15.848 | 15.849 | 15.852 | 15.849 |
| Method | Slope | .938 | .937 | .941 | .980 | .977 |
| | Intercept | 20.823 | 18.588 | 14.020 | -.627 | 13.072 |

Table 3

Raw and Scaled Scores for Two Hypothetical Examinees Across Four Levels of Rekeying

| | Number of Multiple Keys | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 0 | | 4 | | 10 | | 25 | |
| | Raw Score | Scaled Score | Raw Score | Scaled Score | Raw Score | Scaled Score | Raw Score | Scaled Score |
| Examinee A | 120 | 135 | 120 | 132 | 120 | 128 | 120 | 118 |
| Examinee B | 120 | 135 | 124 | 136 | 130 | 138 | 145 | 143 |

Table 4

Distribution of Equators and Total Items Across Content Areas

| | | Content Areas | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| Current Study | % of Equators (SC condition) | 20 | 20 | 15 | 15 | 15 | 15 |
| | % of Total Test Items | 16 | 21 | 16 | 16 | 16 | 16 |
| Klein and Jarjoura | % of Equators (Anchor 1) | 18 | 19 | 4 | 24 | 18 | 17 |
| | % of Equators (Anchor 2) | 12 | 22 | 17 | 16 | 19 | 15 |
| | % of Total Test Items | 20 | 20 | 10 | 20 | 15 | 15 |

## REFERENCES

Dorans, N.J. (1983) Effects on score distributions of deleting an unkeyable item from a test. (Research Rep. No. 83-5) Princeton, N.J.: ETS.

Klein, L.W., & Jarjoura, D. (1985) The importance of content representation for common-item equating with nonrandom groups. Journal of Educational Measurement, 22, 197-206.